Computing the Value-Added of American Postsecondary Institutions[1]

Caroline M. Hoxby[2]

July 2015

ABSTRACT

Computing postsecondary institutions' value-added is an essential step if we are to evaluate the costs and benefits of any policy that affects college-going. For instance, if tax credits and deductions for higher education expenses affect enrollment, the benefits that would offset the costs of these tax expenditures must come from value-added. Similarly, value-added calculations are necessary for evaluating the deductibility of student loan interest, the untaxed nature of many scholarships, tax-preferred education savings accounts, the tax exempt status of most colleges, the deductibility of charitable contributions to colleges, and numerous government spending programs that support higher education. Value-added is also crucial for whether the Treasury will ultimately to recover outstanding student debt. This paper illustrates a method for estimating the value-added of U.S. postsecondary institutions. The key challenge is overcoming vertical selection (some colleges' students are more qualified than others) and horizontal selection (colleges' students may be similarly qualified but differ on geography or family background). We use natural experiments to address selection: quasi-randomization by admissions staff to address vertical selection and quasi-randomization by students to address horizontal selection. We combine the results from the many experiments using paired comparison techniques. We apply the method to comprehensive administrative data on college-going and wage outcomes, and we report policy relevant descriptions of the value-added evidence.

---

[1] The opinions expressed in this paper are those of the author alone and do not necessarily represent the views of the Internal Revenue Service or the U.S. Treasury Department. This work is a component of a larger project examining the effects of federal tax and other expenditures that affect higher education. Selected, de-identified data were accessed through contract TIR-NO-12-P-00378 with the Statistics of Income (SOI) Division at the U.S. Internal Revenue Service. The author gratefully acknowledges the help of Barry W. Johnson and Michael Weber of the Statistics of Income Division, Internal Revenue Service. She also gratefully acknowledges comments from George Bulman, Martin Feldstein, Adam Looney, Jordan Matsudeira, Betsey Stevenson, and Sarah Turner. James Archsmith and Steven E. Stern provided important help with paired comparison methods.

[2] Stanford University and NBER

I. Introduction

There are many reasons why we want to know colleges' value-added--that is, the causal effect on outcomes of attending a specific college. Most obviously, a student may wish to compare a college's value-added to its cost: Can the student expect an improvement in outcomes sufficient to justify her cost of attending? The federal and state governments cannot evaluate their tax and expenditure policies without value-added. For instance, the credit for tuition and fees, which is intended to encourage enrollment, generates a tax expenditure of up to $25 billion a year. Whether this expenditure will be offset by the higher earnings of the college-educated depends on value-added. We cannot evaluate the tax-exempt status of most colleges or the tax deductibility of charitable contributions to colleges without knowing their value-added. We need to understand the value-added of colleges to assess whether their students should qualify for federal student aid or have their tuition subsidized by state appropriations. Federal student debt has reached an unprecedented magnitude, and the Treasury is increasingly involved in collecting debt (through offsets and withholding) and determining who qualifies for income-based repayment. Whether the debt will ultimately be recovered is a question about value-added. Without value-added, we can say little about higher education's potential to affect economic growth and, consequently, the budgetary sustainability of many government programs. Value-added estimates may give us insight into the effects of different college curricula, online versus in-person education, and educational technology. In short, computing value-added is a crucial step in any evaluation of policies that support higher education. It gives us the benefits to weigh against costs.

In a series of studies, we are evaluating federal tax and other programs that support higher education (see Bulman and Hoxby, 2015; Hoxby and Bulman, forthcoming). We are especially interested in comparing the government's return from the various programs, of which there are many. Computing value-added is a building block that supports all the studies.

In theory, we can compute value-added on the basis of any outcome. To name but a few: earnings, employment, public service, inventiveness, marriage, health, child bearing, charitable giving. Earnings are obviously needed for most of the cost-benefit calculations mentioned above, but there is no right or best outcome. For instance, a college that

produces civic-minded public servants may benefit society greatly even if it does not generate the highest value-added when earnings are the measure. Moreover, assigning weights to various outcomes in an attempt to compute some definitive measure of value-added is a not a job for an economist. It is a job for the social planner which is to say (since the planner is only a concept) that it is not any actual person's job. Our own inclination is to compute value-added on as many outcomes as are interesting and measurable. Such a collection of estimates would allow people to assign their own weights based on what they believe colleges should produce.

The challenge we face in estimating the value-added of colleges is that selection is pervasive. There is both a vertical and horizontal nature to selection of students among colleges. Vertical selection (more and less able/prepared/motivated students choosing different colleges) is perhaps the more serious and certainly the better known problem. If not addressed, vertical selection will cause us to overestimate the value-added of colleges whose students are positively selected and to underestimate the value-added of colleges whose students are negatively selected. This leads to the legitimate question that plagues college comparisons: Are the outcomes of students from very selective colleges good because the colleges add value or because their students are so able that they would attain the same outcomes regardless of the college they attended?

However, colleges' student bodies are not only vertically differentiated: they are also horizontally differentiated. That is, they differ on dimensions like geography within a level of ability/preparation/motivation. For instance, suppose that earnings differ across areas of the country for reasons unrelated to higher education. If this is so, two colleges that enroll equally able students and generate equal value-added may have alumni with different earnings. We could easily mistake such earning differences for differences in value-added.

In this environment where selection is pervasive, we can only compute value-added if we first identify plausible "experiments" in which students who are the same end up in different colleges. The primary challenge in this paper is identifying vertical and horizontal experiments--not for a few colleges (since that has been done and is clearly possible) but for virtually all schools. Identifying such experiments is inherently challenging but population data at least makes the exercise feasible. The only schools for

which we do not aspire to compute value-added are those that have inadequate outcome data owing to their very small enrollment or recent creation (their alumni have not sufficiently realized outcomes).

Once we have a full set of experimentally-based differences in outcomes among students who are plausibly the same, we need to combine these results efficiently. This is a problem in paired comparisons and can be dealt with using statistics or algebra. We use statistics. Writing down the paired comparison problem is straightforward, but implementing it on the scale we do is logistically challenging and possibly unprecedented.

Among the potential outcomes we can use to record value-added, earnings are the most obvious because they are needed for almost any tax or spending policy analysis. Moreover, earnings are continuous, inherently cardinal, and distributed in a manner we understand. This makes earnings a much easier outcome with which to work, as a statistical matter, than outcomes that are ordinal, categorical, or distributed in a less well-understood way: occupation, sector, marriage, children, etc. Since this paper is primarily methodological, we do not wish to involve ourselves in unnecessary problems. Therefore, we only compute earnings-based measures of value-added in this paper. However, our method could be applied to any outcome with the caveat that estimation would be more burdensome computationally with a categorical outcome.[3]

To see intuitively where our experiments come from, recognize that the college choice process is imperfectly informed. It is the imperfect informedness of this process and our understanding of these imperfections that generate the experiments.

For our vertical experiments, we exploit the fact that admissions processes, even the most careful ones, necessarily choose a subset of students in an arbitrary way. These students are often described as being "on the bubble" because in an initial triage based on standard qualifications such as test scores, they are in a range where admission and rejection are equally likely. (The evocative phrase "on the bubble" is an American idiom described in the footnote.[4]) A selective school's bubble range of test scores is identifiable

---

[3] We are simply pointing out that estimating a multinomial probit or logit model is more burdensome, computationally, than estimating a linear model.

[4] The Cambridge Dictionary of American Idioms (2003) defines "on the bubble" as "equally likely to experience either of two results." The phrase is "based on the idea that something on the

because, above it, a high percentage of students are admitted and admission rates decline slowly in test scores. Below the bubble range, almost all students are rejected: a student must demonstrate some exceptional quality in order to be admitted. In the bubble range, admissions probabilities are such that staff appear to be flipping coins. Of course, staff are not actually flipping coins. Rather, they are influenced by minor considerations that could not plausibly have a major effect on long-term outcomes except through the admissions decision. For instance, an admissions staff member may happen to prefer the extracurricular area in which the student expresses an interest. Intuitively, we address vertical selection by treating each school's bubble as a range in which admission is randomized. Later we are much more precise about how we identify each selective college's bubble and how we use it. We also discuss nonselective colleges that have no bubble because they enroll any student who registers.

For our horizontal experiments, we exploit the fact that students have an imperfect understanding of the college options available to them. Because they do not perfectly understand which colleges have the greatest value-added for them, they may regard colleges as similar that in fact differ in value-added and other qualities. On the one hand, this is not at all surprising: reliable value-added information is, in fact, not available to students. Comparability is also obscured because colleges' net prices have traditionally been difficult to learn.[5] In any case, we observe that students treat certain colleges as an "indifference set": the same students apply to them and the colleges end up with student bodies that are the same on vertical selection. That is, their students have very similar assessment scores, grades, and family backgrounds because the students are randomly choosing among them. While few students actually roll a die to choose a college from their indifference set, they choose on grounds that are trivial determinants of outcomes: the appeal of particular buildings, the weather on the day they visited, the off-hand suggestion

---

surface of a bubble is as likely to roll in one direction as in another." Examples of usage include: "These are the players on the bubble, the ones who are not sure if they have made the team" (Cambridge Dictionary 2003) and "Some states will vote for the Democrats, and some are likely to vote for the Republicans, but Arizona is on the bubble." (American Heritage Dictionary 2011).

[5] Even today, with net price calculators available online, families can get only an approximate sense of colleges' net prices.

of an acquaintance. Intuitively, we address horizontal selection by identifying indifference sets of colleges and treating students' choice among them as random. Later we are much more precise but suffice it to say that our indifference sets are tightly defined and based both on applicants and enrollees.

Our vertical and horizontal experiments provide us with different types of information. The vertical experiments help us understand differences in value-added between colleges that may be at different selectivity levels but whose students overlap because they are on the bubble at the more selective college. The horizontal experiments help us understand the differences in value-added among schools whose students bodies exhibit very similar ability, motivation, and preparation. The two types of experiments generate different information, a point discussed below.

The contribution of this paper is four-fold. Most importantly, we illustrate a method for estimating U.S. colleges' value-added. We believe that this method is as reliable and plausible as any method available to us. It is certainly much more reliable than the extremely crude methods employed by popular websites, media, and many organizations charged with policy-making in higher education. Their methods often do not address selection at all. We would not, however, argue that our method is perfect. Unless we conduct vast randomized controlled trials in which we randomly assign students to colleges, we cannot obtain purely experimental estimates of value-added. Such trials would be grossly unethical on numerous grounds so we must attempt to generate estimates using "natural experiments" such as those we describe.

Second, we apply the method to nearly comprehensive, accurate, administrative data on college-going and outcomes. The accuracy and density of our data are what allow us to use the methods we do: they are crucial for identifying colleges' bubbles and indifference sets. They are also important for providing an accurate picture of outcomes associated with each college.

Third, we use paired comparison techniques to build the entire scale of value-added. The evidence generated by the experiments is essentially of the paired type: outcomes for the same type of student at college A versus college B. Yet, we want to construct value-added across all colleges. This is the paired comparisons problem. While we do not consider this paper to be an innovation in statistics or ranking theory, we implement

paired comparison techniques on an unprecedented scale. It is noteworthy that the paired comparison method is not at all equivalent to controlling for observable measures of students' ability using a regression. That method would impose numerous distributional assumptions about unobservable variables and numerous functional form assumptions about how observable measures of aptitude determine outcomes. Our method instead treats each experiment as an experiment in which we assume only that equal students experience different colleges at random. Having defined the experiments, the rest of the method is as transparent as we can make it. Moreover, we are open to refining the rules that define the experiments. Such refinements are a form of robustness testing rather than a change in the method itself.

Finally, we obtain value-added estimates for the U.S. postsecondary institutions that enroll the vast majority of undergraduate students.

It is worthwhile saying what this paper does not do. First, it does not list the value-added of each of the thousands of U.S. postsecondary institutions. It only reports value-added estimates for clusters of institutions where the clusters are defined in objective ways. The cluster-level estimates answer policy questions but focus the reader on methodology in this primarily methodological paper. Second, this paper does not attempt to make rate of return estimates which would require an examination not only of schools' differences in value-added but also causal differences in the educational costs they generate. Rate of return estimates are a natural next exercise discussed near the end of the paper. Third, in our experiments, a student chooses college A or college B. Many consequences may endogenously follow. For instance, the student may persist in college A while she would have dropped out at college B. Or, college A may induce the student to enter graduate school while college B would not. We can explore persistence, further education, and other endogenous variables as outcomes affected by the same experiments, but our method uses experiments in initial college-going. It does not have experiments that separately identify the effects of persistence, separately identify the effects of further education, and so on. Fourth, this paper does not attempt to estimate value-added within colleges by major, program, or otherwise. Indeed, until the final sections of the paper, we assume that each college's value-added is constant across its students and programs. However, we discuss these issues later. Fifth, for purely practical reasons, we do not

attempt to estimate value-added among students who are not of traditional college-going age. They are for future work.

The remainder of the paper proceeds as follows. In section II, we explain why value-added estimates are important and describe desirable attributes for such estimates. In section III, we explain why value-added estimates have previously not existed for the vast majority of postsecondary schools. We also discuss previous literature that clarifies the logic of our method. Section IV describes the data. In sections V and VI, respectively, we discuss our vertical and horizontal experiments. In section VII, we draw upon paired comparison techniques to combine the results of our experiments efficiently into a comprehensive value-added scale. Section VII briefly summarizes the method and extends it to students who consider only nonselective schools. Section IX contains our main results, the value-added estimates. We discuss extensions, robustness checks, and outstanding issues in section X.

II. Estimates of the Value-added of Postsecondary Institutions

A. The Crucial Nature of Value-added Estimates

Individuals, societies, and governments need estimates of the causal or selection-purged value-added of each postsecondary institution. When considering a specific postsecondary school (versus not attending at all or versus attending another postsecondary school), it is fairly easy to estimate the costs, even the opportunity costs (lost earnings etc.). Thus, the crucial element that is often missing for deciding whether a postsecondary school is a good or bad investment is the value-added.

For instance, if the federal government is considering whether a program, such as a tax credit for tuition and fees, can support itself fiscally, it must first determine how the program affects each school's enrollment and then convert the effects into lifetime taxable earnings and other measures of ability-to-pay and dependency that are relevant to the federal budget.

More broadly, value-added is crucial to any policy or problem in which college potentially affected outcomes. The classic human capital investment problem dictates that a person should attend postsecondary school j if

(1) $$U(Y_{ij}, c_{ij}) > U(Y_{ij'}, c_{ij'})$$

for all schools j' not equal to j. U is lifetime utility. i indexes the person. $Y_{ij}=(Y_{ij}^{1}, Y_{ij}^{2}, ..., Y_{ij}^{k})$ is a vector of lifetime outcomes that affect utility and that are, in turn, causally affected by postsecondary school; and c is the present value of the direct cost of the school (tuition and fees and any financing costs associated with paying them).[6] The elements of Y may be lifetime earnings, inventiveness, ability to contribute to society, marriage, health, and so on. Note that j=0 represents the outside option of attending no postsecondary school at all.

Some utility functions are such that we can rewrite (1) to emphasize that an individual's decision to attend a school should be affected by its added value vis-a-vis other schools on each outcome: $Y_{ij}^{1} - Y_{ij'}^{1}$, $Y_{ij}^{2} - Y_{ij'}^{2}$, ...

In addition to the private benefits that an individual may enjoy from her education, society may enjoy externalities. For instance, postsecondary education may make people better participants in civic life or better at insuring themselves against risk. These social benefits of education do not fundamentally change the nature of the investment problem but simply mean that social welfare, social outcomes, and social costs should be used rather than their individual counterparts.[7] That is, society would like to induce a person to attend a postsecondary school if

(2) $$W(\check{Y}_{ij}, \check{c}_{ij}) > W(\check{Y}_{ij'}, \check{c}_{ij'})$$

where W is social welfare; $\check{Y}_{ij}=(Y_{ij}^{1}, Y_{ij}^{2}, ..., Y_{ij}^{k}, Y_{ij}^{k+1}, ..., Y_{ij}^{K})$ is the more inclusive vector of outcomes that affect social welfare and that are, in turn, causally affected by postsecondary school; and č is the present value of the social direct cost of the school which may include costs borne by taxpayers or philanthropists.

In short, estimates of value-added are crucial for evaluating many policies and making optimal investments in higher education.

B. Desirable Attributes of a Method for Estimating Value-added

Given the importance of the estimates, a method for computing postsecondary schools'

---

[6] For simplicity, the dynamics of the problem are subsumed by expressing everything in lifetime terms.

[7] Also, the social discount rate should be used rather than the individual discount rate.

value-added should display several attributes.[8]

i. Attention to Causality. It is crucial that value-added estimates be causal, not the differences in outcomes among schools that include the effects of selection. Selection-included effects are not useful for most policy-making or for individuals deciding where and whether to attend college. Indeed, selected-included effects can so grossly misrepresent the causal effects that an entity that encourages individuals to rely on selection-included effects may harm individuals and the nation's fiscal situation by inducing people to engage in education whose value-added will not justify the costs. Not only do individuals waste their time and income (relative to more productive uses), they may end up with un-repayable student debt or they may linger in jobs with wages below the threshold where repayment begins.

The importance of causal estimates appears often to be missed because many thought leaders quote selection-included differences in outcomes in an effort to induce individuals to enroll in postsecondary school. Organizations that students view as authoritative routinely publish selection-included estimates without warnings that they are not causal and should not be used for decision-making. The residual claimants of losses due to mistaken decision-making are the students themselves but also (given the importance of government tax expenditures, aid, and student loans) the Treasury and its state counterparts. Therefore, governments have incentives to inform students about casual effects: the government budget internalizes some of the consequences of mistakes.

As with medical treatments, it is probably impossible to generate estimates of schools' causal effects that are wholly free from selection bias. However, one should try one's best if the estimates are intended for a use that requires causal effects.

ii. Accuracy. Value-added estimates should accurately reflect the effects of postsecondary schools. Unbiased but very noisy estimates are undesirable.

iii. Comprehensibility. The method should be sufficiently comprehensible that an intelligent lay person could understand it at least on an intuitive level. In particular, readers should understand the natural experiments well enough to see what drives the estimates. Also, when choosing among statistical and/or mathematical methods, weight

[8] We credit Harville (2003) with listing some of these attributes although we describe them differently because our application is colleges' value-added while his is seeding basketball teams.

should be given to comprehensibility.

iv. Using What We Know about College Choice. Our value-added method should incorporate what we know about the college choice process. For instance, we know that certain data, most obviously scores on college assessments like the SAT or ACT, are routinely used by schools in the admissions process. Similarly, we know that other factors, most obviously geography, strongly influence students' choices of nonselective and less selective schools. (Geography is much less important for high achievers likely to be admitted by selective institutions that compete for students nationally or at least state-wide.)

v. Non-Manipulability. The method should not rely on data that are easily manipulable by schools. Most obviously, it should not rely on schools' self-reported selectivity.

Thus, from now on, "selectivity" refers to the assessment scores of a college's enrolled students, which we measure rather than relying on colleges' self-reports. There are only two reasonably defensible measures of selectivity: revealed preference and enrolled students' scores. They are very highly correlated so there is not much to choose between them in practice. Although admissions rates and matriculation (or yield) rates are thought by many people to be revealed preference measures, they are not in fact reasonable substitutes for properly constructed measures. They are not even more than trivially correlated with proper measures (Avery, Glickman, Hoxby, and Metrick 2013). Moreover, colleges can manipulate their admissions and matriculation rates to make themselves appear to be more selective than they are. Assessment scores and properly constructed revealed preference are not manipulable.

III. What We Derive From Previous Value-added Estimates

A. Why Credible Value-added Estimates Have Previously Not Existed

Prior to this paper, credible estimates of value-added have previously not existed for the vast majority of U.S. postsecondary institutions. There are important exceptions to this rule, as emphasized below, but estimates have been lacking. Why, if such estimates are so crucial, have they not existed?

The first reason is that data that link college attendance to post-college outcomes, such as earnings, have been lacking. While such data exist in federally-supported studies

such as the National Longitudinal Survey of Youth and the National Educational Longitudinal Study, longitudinal studies like these have samples that are far too small to estimate value-added for individual institutions. Many institutions would not be represented by even a single student. Those that would be represented would typically be associated with only a handful of students. As an alternative to federally-supported surveys, some firms have attempted to gather outcome data for institutions' students through crowd-sourcing or commercial samples. Such data are much less appropriate than the those from the federal surveys since the firms' samples are not only very small but suffer from self-selection: the people who voluntarily report their outcomes can be egregiously non-representative. Also, there is no reason to believe that their reports are accurate.

The second reason why value-added estimates have not existed is that it is nearly impossible to address selection problems without comprehensive administrative data. Sample- or survey-based data are insufficient for the methods outlined in this paper. Indeed, addressing the selection problems is what is demanding in terms of data and methods. The simple linking of outcomes to colleges involves no real difficulty if representative, accurate data are available.

The final reason why such estimates have not existed is subtle but boils down to researchers failing to identify and/or efficiently use experiments that credibly address the selection problems. This failure is probably the indirect result of the aforementioned data paucity. Sensible researchers do not invest much energy into developing methods for which there are no appropriate data.

B. Previous Methods that Credibly Address Vertical Selection

As mentioned above, there are a small number of studies that have generated credible estimates of value-added--usually for a single institution or a small set of institutions. These exceptional studies rely on regression discontinuity (RD) methods. For example, Hoekstra (2009) investigates a state's flagship public university that admits students using a fairly strict score cut-off on college assessment exams. Using fuzzy RD methods, he compares students who are just above the cut-off (usually admitted) to students just below the cut-off (usually rejected). He thereby generates a credible treatment-on-the-treated estimate of the university's value-added relative to the pool of institutions that the rejected

students attend. (The pool is dominated by public colleges that are in the same state but that are less selective than the flagship university.)

Other studies that use RD methods to obtain value-added estimates include Saavedra (2008), Kaufmann, Messner, and Solis (2012), Cohodes and Goodman (2012), Hastings, Neilson, and Zimmerman (2013), Zimmerman (2014), and Goodman, Hurwitz, and Smith (2015). Among these, the U.S.-based studies usually focus on one or a small number of institutions because most U.S. schools do not employ strict score cut-offs. It is worth noting that only some of the RD studies generate value-added for post-college outcomes such as earnings. Others generate value-added only for outcomes that appear in the institution's own data (degree completion, for instance).

RD methods depend on continuity assumptions for identification. In contrast, this paper's vertical experiments depend on a randomization assumption. These identifying assumptions are not equivalent. Nevertheless, there is analogous logic because both methods are based on the observation that a subset of students are on the bubble in a selective college's admission process. Staff are forced to choose among the bubble students in an arbitrary way. In schools to which RD applies, the arbitrariness is generated by a cut-off that is so sharp that being on one or the other side of it is as good as random. In the vertical experiments, the arbitrariness is generated by some factor so minor that it is as good as random. That factor could any number of things that matter to a particular admissions staff (including but not limited to a score cut-off).[9]

C. Previous Methods that Could Credibly Address Horizontal Selection

Dale and Krueger (2002) introduce a method that, applied appropriately, could address horizontal selection. It is related to our method but they apply it differently so it is best to discuss it in the abstract.

Suppose two students are admitted to the same set of postsecondary institutions: schools A, D, K, M, and P. The fact that their admitted school portfolio is identical suggests that, first, the students are interested in the same sort of schools (since they evidently applied to them) and, second, the students' qualifications for college are similar.

[9] In fact, the vertical experiment will generate results very similar to fuzzy RD for any school that actually admits students based on a fairly sharp score cut-off. In such cases, we find a narrow bubble range that is roughly equivalent to the bandwidth in RD.

That is, vertical selection on the basis of interests, qualifications, and motivation is very plausibly the same for the two students.

Suppose that the two students are indifferent between schools A, D, K, M, and P. Then each student will choose the school in which he enrolls by rolling a die or using some equally random process. We can then compare the two students' outcomes to obtain an estimate of the value-added of one of the schools versus another of the schools. If there are many students with the same portfolio, we can obtain value-added for A versus D, D versus K, K versus M, M versus P, and all other pairs.

Since this method relies on students (not admissions staff) randomizing, it is useful for estimating value-added in the face of horizontal selection. It is very important to note that this method is only credible if, ex post, it turns out that schools A, D, K, M, and P are actually equally selective on a vertical basis. If they are not, then students could not possibly be randomizing among the schools and there is no credible experiment. Thus, this method can only be applied when combined with verification that the schools are equally selective vertically.

To make this concrete, suppose that the two students are indifferent between schools A, D, K, and M but that school P is a less selective institution included in the portfolio only as a "safety school." Then if one of the two students chooses school P, he could not be randomizing. He must be deliberately choosing P for a reason unknown to the researcher--perhaps his parents are important donors to school P and he expects special treatment as a result.

What matters is that the following cannot simultaneously hold: (i) students with the same interests and qualifications randomize among school P and the other schools in its indifference set; (ii) school P is less selective in equilibrium. Either (i) holds, in which case the schools exhibit the same vertical selectivity, or (ii) holds, in which case the students are not randomizing but revealing that they differ in some important way not known to the researcher but which nevertheless violates the assumptions of the exercise. In other words, the method based on students randomizing cannot be used to address vertical selection because there are no mutually consistent conditions under which vertical selection would be remedied. The method can only be used to address horizontal selection.

Dale and Krueger (2002) attempt to use the method to address vertical selection.

Since the identifying assumptions are inconsistent in such an application, we have not described their study in detail. The aforegoing discussion is not intended to criticize their study but simply to clarify that the method they propose is only useful for addressing horizontal selection.

This paper addresses horizontal selection using a enriched version of the method based on students randomizing. Crucially, our method builds in verification that the schools in an indifference set are actually equally selective vertically.

IV. Data and Types

A. Data

We use administrative data on college assessment scores, score sending, postsecondary enrollment, and 2014 earnings from wages and salaries for people in the high school graduating classes of 1999 through 2003 who were aged 29 through 34 in 2014.[10] Score data are from The College Board, enrollment data from the National Student Clearinghouse, and earnings from de-identified Form W-2 data.[11] Prior to use, all the data are not just de-identified but "collapsed" or aggregated to a group level. Our method only requires mean outcomes by type-treatment group where a type is students from the same cohort who have the same scores and applied to the same postsecondary school. A treatment is enrollment in a particular postsecondary school. Because our method requires only this group-level data, not individual-level data, only group-level data were obtained for analysis.[12] Note, however, that we continue to use words like "students" or "people" instead of "types." We do this simply because it is awkward to say that "types" engage in some behavior. However, readers should keep in mind that the method is actually applied to group-level, not individual, data.

Chetty, Friedman, and Rockoff (2014) establish that earnings at ages 29 to 34 are sufficiently informative that older-age data does help much in predicting a person's lifetime

[10] That is, we employ data on students who graduated from high school at age 18 or 19, which are the dominant ages at high school graduation in the U.S.

[11] ACT data are used in a limited capacity as noted below.

[12] That is, data queries returned only group-level mean outcomes and group codes.

earnings. Since we would be forced to estimate value-added for less recent students if we required older-age data, we believe the cohorts we use are a reasonable set.

When applying to a selective college, students nearly always have their scores sent. (Applicants also often send their scores to nonselective colleges for diagnostic purposes or avoidance of remedial courses.) Score sending is therefore a widely used proxy for a student's applying. It has been shown to be an accurate one for selective colleges that normally use scores in admissions (Hoxby and Turner 2013). Hereafter, we refer to a score sender as an "applicant" and score sending as "applying." This allows us to avoid sentences that are so awkward that they can easily be misinterpreted.

B. Student Types

A few points on the types are worth making. First, each type is uniform not only in terms of tested aptitude but also in terms of interest in a particular postsecondary school since that interest is revealed by applying.

Second, we could define the types more finely or coarsely. For instance, we could use family income to define finer types. This would be useful if admissions staff treat same-scoring applicants differently depending on their ability to pay. A small share of schools (i) do not conduct need-blind admissions and (ii) commit to fully meeting a student's financial need. Such schools are known to consider a student's need because they could otherwise outrun their aid budgets. However, most schools do not fit these two criteria simultaneously. The most selective U.S. schools conduct need-blind admissions. Less selective schools almost never pre-commit to fully meeting need. They offer admission and describe the financial aid available. If the student cannot afford the school, he simply does not enroll. Nevertheless, we remain concerned about the possible need to use family income to refine the types. Later in the paper, we use high schools in our definition of types partly to ensure that students who are regarded as the same actually have similar socio-economic circumstances. In any case, family income is an issue to which we return.

We could also use high school grades to define finer types. However, this might not be helpful because college admissions staff treat grades differently depending on the high school that issued them: different high schools have different grading standards. Other refinements that we considered seemed likely to introduce more noise than information. In any case, we are open to making the types finer or coarser but view such alterations as

robustness checks rather than a choice among methods.

Third, there are many students who, despite taking assessments (so that they have scores), do not send scores to any school because they only consider open-enrollment schools. We describe later how we organize such students into types. For now, it is useful to focus on students who apply to some school. They make the method clear and we can later extend it easily.

V. On-the-Bubble Experiments that Address Vertical Selection

To address vertical selection, we make use of the fact that selective colleges have test score ranges where applicants are on the bubble. Within the bubble range, students are fairly likely to be either admitted or rejected and the decision may turn on a minor factor. Essentially, we make use of unintentional randomization conducted by admissions staff.

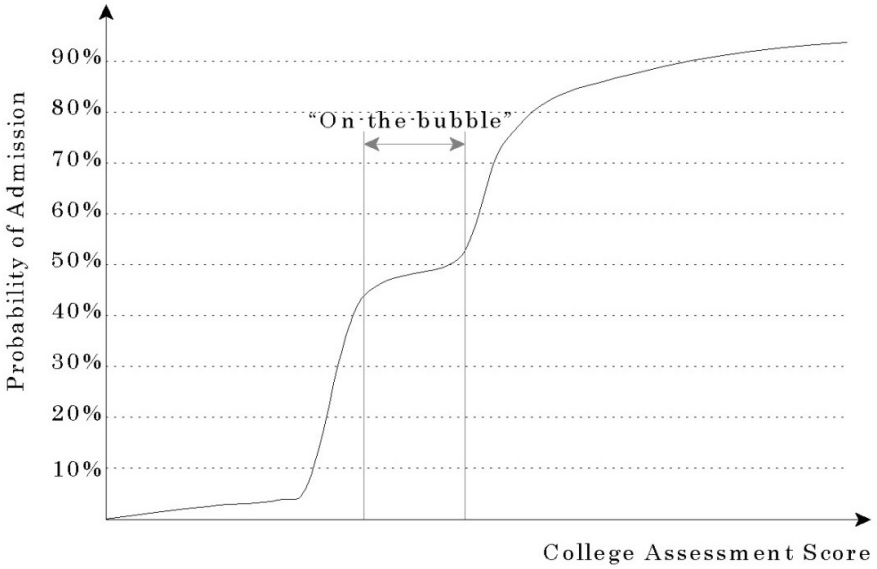
The existence of bubble ranges is an observation about what we actually see in the data. It is not an assumption. Moreover, we do not see bubbles on other criteria such as high school grades, distance from the college, and so on. These observations suggest that, probably because they are quantitative measures, scores play a dominant role in triage.

It may be useful to know that bubble ranges are used not just in admissions but many other selection procedures: triage of job applicants, loan applicants, potential venture capital investments, and so on. Bubble ranges exist because selectors need to conduct triage in order to focus their effort on marginal cases and because, in triage, selectors tend to categorize applicants for later, more holistic scrutiny.

Bubble ranges generate distinctive profiles in college's admissions and enrollment rates, as shown in Figure 1. (Figure 1 is based on admissions data from several colleges that are very similar in their selectivity.) In the figure, admissions probabilities are high above the bubble, low below the bubble, and intermediate in the bubble.

To identify each college's bubble range, we could fit (separately for math and verbal scores) admission outcomes to test scores using kernel-weighted local cubic polynomial regression. We would then choose the range where the smoothed values indicate that the admission probability is between 40 and 60 percent. (We can vary these boundary probabilities as a robustness check.) We would do this for each U.S. institution except as noted below.

Figure 1
Probability of admission by score, illustrating an on-the-bubble range



In fact, we have enrollment but not admission outcomes so we adapt this procedure slightly. We fit enrollment outcomes to test scores using kernel-weighted local cubic polynomial regression and choose the range where the smoothed values indicate that the enrollment probability is between 40 and 60 percent of the maximum enrollment probability.[13] (We have experimented with these boundary probabilities as a robustness check.)

We do not search for bubble ranges at schools that have open enrollment policies or where fewer than 75 percent of applicants submit scores. These schools do not practice vertical selection. Furthermore, if our bubble criteria are not fulfilled by any range that covers at least 10 percent of the school's test score range, we describe the school as having no bubble. In practice, no bubble schools tend to be nonselective.

We find that bubble ranges are below the score of the median enrolled student but

[13] Using the smoothed values, we take the maximum probability over a range that includes 20 percentiles of the school's distribution of test scores. As an empirical matter, this range is usually around the school's median test score. We also check that applicants in the bubble range who do not enroll usually attend equally or less selective schools. The check ensures that we do not identify ranges where schools reject students because they are "overqualified" and therefore unlikely to matriculate. In practice, such strategic rejection appears to be so rare that the check is unnecessary. We use ACT data to ensure that the probabilities and percentages are correct for schools where many applicants submit the ACT.

usually not among the very lowest scores (the bottom 5 to 10 percent) among enrolled students. The bottom 5 to 10 percent of students who enroll frequently have scores so much below those in the bubble range that they must have been admitted on peculiar grounds (such as athletic recruitment) so that their scores played a fundamentally different role in initial triage.[14]

Having identified each school's bubble range, we treat the applicants in this range as randomly admitted or rejected. Thus, we need only compute the difference in outcomes for each observed treatment pair for each type. For instance, if we considering applicants who are on the bubble at school A, we compute the difference in outcomes for those who attend school A versus school B, and we do this separately for each type of student. We also compute the difference for school A versus school C, school A versus school D, and so on until we exhaust all of the observed college pairs and student types.[15]

Although onerous, this procedure is essentially simple and intuitive. This simplicity is the direct result of identifying natural experiments. At the end of the procedure, we have the results of all pairwise college-versus-college experiments for students who were on the bubble at some selective school.

VI. Indifference Set Experiments that Address Horizontal Selection

To address horizontal selection, we make use of the fact that students select colleges fairly at random within their indifference sets. This series of natural experiments allows us to identify plausibly causal differences in value-added among equally selective colleges.

Suppose we know that a student applies to colleges A, D, K, M and P. How do we know whether these schools are an indifference set--that the student is randomizing among them? Recall that schools in the same indifference set must not only attract the same applicants but must end up with the same selectivity. Thus, the first step in the indifference set experiments is identifying, for each college, all other colleges with the same selectivity. The second step is identifying the students who randomize among this college

[14] Put another way, many schools exhibit a long but very thin left-hand tail in test scores among admitted students.

[15] Because we use type-level data, we keep track of the number of students represented in each comparison.

and the others in its indifference set.

To implement the first step, we compute the empirical test score percentiles of each college's enrolled students.[16] We then define each college's potential equals as those that have the same 25th and 75th percentile scores on the both the math and verbal tests. We do not require exact score matches at each of these four percentiles but allow a cushion of plus or minus 3 national percentiles. (As robustness checks, we have experimented with using more and different percentiles--for instance, the 20th, 50th, and 80th. We have also experimented with different cushions such as plus or minus 2 or 5 national percentiles.)

To implement the second step, we consider each possible pair of colleges that are potential equals. We then identify all those students who apply to both schools in the pair and who are very likely to be admitted to both colleges conditional on applying. We impose the second condition because the indifference set-based experiments should be based on students who get to choose between the schools, not those whose choice is made for them by a rejection letter. We say that a student is very likely to be admitted to a college if his or her test scores put him between the 65th and 80th percentiles on that school's enrollment students' score distribution.[17] (As robustness checks, we have experimented with different percentile ranges--for instance, the 60th through 75th and the 70th through 85th. The results turn not to be sensitive to these changes but we are open to all reasonable alternatives.)

We have now identified all of the possible horizontal experiments: occasions where students get to choose between two equally selective colleges in which they have shown an interest. We treat these choices as random and, thus, need only compute the difference in outcomes for each observed treatment pair for each type. For instance, if we considering students choosing between equally selective schools A and B, we compute the difference in

---

[16] We compute the empirical percentiles ourselves since some colleges publicize percentiles that are inaccurate. We use ACT data to ensure that the percentiles are correct for schools where many applicants submit the ACT.

[17] If we had each student's admittances, we would probably use those at this point. However, what we have done is nearly equivalent in the data because (i) we have selected score ranges where admittance rates are very high and (ii) we are conditioning on the two colleges being equally selective so the applicant will strike each college the same way vis-a-vis its enrolled student body.

outcomes for those who attend school A versus school B, and we do this separately for each student type. We do this until we exhaust all of the relevant observed pairs and student types.[18]

This procedure is onerous but intuitive because it depends transparently on our identifying the horizontal natural experiments. At the end of this procedure, we have the results of all pairwise college-versus-college experiments for students choosing between equally selective colleges in which they have shown an interest.

VII. Using Paired Comparison Methods to Efficiently Combine the Results of all the Vertical and Horizontal Experiments

Upon completing the procedures described in the two previous sections, we have the results of all the observed pairwise college-versus-college experiments, both vertical (admissions staff randomizing among students on the bubble) and horizontal (students randomizing among equally selective schools). In this section, we explain how we use paired comparison techniques to combine all of these college-versus-college or "head-to-head" experimental results into a value-added scale.

A. Why Combine the Results of the Head-to-Head Experiments?

It is worthwhile explaining why this exercise is valuable because, even before being combined, the head-to-head results are useful for answering certain questions. For instance, suppose a student knew ex ante that he was only interested in colleges A and B. He might be content with learning their head-to-head results as indication of their respective value-added.

However, there are a few reasons to combine the head-to-head results efficiently. First, our goal is to construct a value-added scale that covers nearly all institutions. This is important because most students should not limit themselves ex ante to comparing a few schools that happen to have numerous head-to-head experiments. Yet, they cannot conduct a wider search without the full value-added scale. By combining the head-to-head results efficiently, we "connect" schools that have infrequent head-to-head experiments

[18] Because we use type-level data, we keep track of the number of students represented in each comparison.

because they are horizontally--for instance, geographically--differentiated. The connection allows students to compare colleges broadly, not just compare a few local schools.

The second reason to combine the head-to-head results is that nearly all policy questions require a scale. For instance, the U.S. Department of Education must presumably set financial aid policies that treat all similar college choices similarly. (That is, policies should be seen to be non-discriminatory.) However, without a value-added scale, the department could not judge similarity across college choices.

The third and most important reason for combining the head-to-head experiments is that we are discarding a great deal of valid information if we do not do it. Combining the experiments is very informative because each experiment provides an implicit check or cross-validation of others. To take the simplest example, suppose colleges A and B are equally selective and have very similar bubble ranges. Suppose further that the results of their head-to-head horizontal experiments suggest that they have equal value-added. If both have head-to-head vertical experiments versus college C, these vertical experiments should confirm that A and B have equal value-added. If we fail to use the information from the A-versus-C and B-versus-C experiments, we are throwing information away.

The fourth and final reason for combining the experiments is related to the fact that the vertical and horizontal experiments are local to different sorts of students: on-the-bubble versus highly-likely-to-be-admitted. We return to this point below. For now, we assume that each school's value-added is the same across the types of students it enrolls.

B. Applying Paired Comparison Methods to the Value-Added Problem

Combining all of the head-to-head results to derive a value-added scale is a problem for paired comparisons techniques. These techniques are widely used across an array of applications where agents compare multiple alternatives but do not rate or rank all alternatives simultaneously. (What matters is not whether a pair is considered--an agent might compare several alternatives--but whether we must construct a some universal scale from non-universal comparisons.)[19]

[19] Paired comparison problems often arise in sports and certain games like chess, and some readers may find the analogy instructive. (Much of the language used in paired comparison methods is based on sports: "head-to-head.") The closest analogous problem in sports is using point spreads from all competitors' past head-to-head meetings to predict the point spread that would occur if any pair of competitors were to meet. See Chapter 9 of Langville and Meyer (2012), Stern (2011), and

Consider a vector of the true values of colleges j = 1 through J on some outcome such as earnings:[20]

$$\boldsymbol{V}=\begin{pmatrix} v_1 \\ v_2 \\ \vdots \\ v_J \end{pmatrix} \quad . \tag{3}$$

If there were no sources of noise in the outcome, each value difference v_j-$v_{j'}$ would equal the difference in outcomes Y_j-$Y_{j'}$ that students would attain if experimentally assigned to enroll in the "treatment" school j instead of the "non-treatment" school j'. We could gather up all of the value differences in a value-differential matrix V

$$\boldsymbol{V}=\begin{pmatrix} 0 & v_1-v_2 & \dots & v_1-v_J \\ v_2-v_1 & 0 & \dots & v_2-v_J \\ \vdots & & & \\ v_J-v_1 & v_J-v_2 & \dots & 0 \end{pmatrix} \quad . \tag{4}$$

and gather all of the outcome differences in an outcome-differential matrix

$$\boldsymbol{K}=\begin{pmatrix} 0 & Y_1-Y_2 & \dots & Y_1-Y_J \\ Y_2-Y_1 & 0 & \dots & Y_2-Y_J \\ \vdots & & & \\ Y_n-Y_1 & Y_n-Y_2 & \dots & 0 \end{pmatrix} \quad . \tag{5}$$

Harville (2003). However, the analogy to sports can be a distraction because many of the problems that arise in paired comparisons in sports do not occur in the college value-added problem. For instance, the number of head-to-head experiments in sports is often tiny (competitors may encounter one another only once each year) while two colleges may have dozens, hundreds, or even thousands of experiments in each cohort of students. Outcome variables (football scores, for instance) are often very lumpy in sports, but outcomes like earnings are continuous. Winning or losing a head-to-head competition matters in sports even if the margin is tiny. There is no similar importance to tiny value-added margins. In short, to the extent that readers find the analogy to sports helpful or clarifying, they may use it. However, a reader who finds himself trying to find connections to paired comparison methods for a specific sport is more likely confuse himself than clarify matters.

[20] Since we consider only one outcome from now on, we drop the superscript on the outcomes for notational convenience.

Of course, there are sources of noise in real-world outcomes, so each experimental result contains an error. The best we can do is choose estimates of each college's value-added that minimize the difference or "error" between the two matrices. That is, we want to find the J×1 vector θ that minimizes

$$f(\theta) = \|K - V(\theta)\|^2 \tag{6}$$

for some matrix norm.[21] If the norm we choose is minimizing the sum of squares of the errors, we end up with the straightforward regression:

$$Y_{iij'} - Y_{ij'j} = D_j\theta - D_{j'}\theta + (\varepsilon_{ij} - \varepsilon_{ij'}) = (D_j - D_{j'})\theta + u \tag{7}$$

where the left hand side is a vector of experimental results with $N \times J^2$ rows (one for each experiment which is indexed by the type i, the treatment/enrollment school j, and the non-treatment/non-enrollment school j'). The matrix D_j has $N \times J^2$ rows and J columns and is made up 1s and 0s such that the cell in column j is equal to one whenever the relevant school is the treatment/enrollment school in the experiment and equal to zero otherwise. The matrix $D_{j'}$ also has $N \times J^2$ rows and J columns and is made up 1s and 0s such that the cell in column j' is equal to one whenever the relevant school is the non-treatment/non-enrollment school in the experiment and zero otherwise.

This regression has very simple intuition. It is merely finding the school fixed-effects that best explain the experimental results. The fixed effects estimates are the estimates of the value of each school.

Notice that we have imposed very little to get to this point. We have assumed that each school's value is fixed across all types of students. (This is not a trivial assumption and is one to which we return later.) We have decided to minimize the sum of squared errors rather than some other norm. This is something that could be explored in a robustness check but is standard in both the statistical and algebraic literature on paired comparisons.

One caveat: Because our data are at the type-by-treatment level and not the individual student level, some experimental results are based on more students than others. Thus, we run weighted least squares regressions.

---

[21] Langville and Mayer (2012) illustrate a purely algebraic implementation but note that paired comparison work is increasingly implemented by various regression methods.

C. Standard Errors and Predictions

When we estimate the aforementioned regression, we generate standard errors on the value-added estimates. Since the data are comprehensive administrative data, not sample data, the standard errors should not be given a sampling interpretation. We suggest an interpretation based on counterfactual states of the world in which the student's "hand trembled" and she picked college A instead of equally selective college B. Or, the counterfactual might be that the admissions staff member's "hand trembled" and an on-the-bubble student was rejected rather than accepted. Put another way, the standard errors are informative for a student who wants to know how her outcomes would have changed had her college choice gone another way, all else equal.

We suspect, however, that the main concern about value-added estimates is that people wish to use them for predictive purposes when they are--necessarily--based on past cohorts. If the macro economy, technology, social environment, or colleges have changed a great deal between those cohorts and the current one, the value-added estimates will not reflect these changes. Furthermore, the standard errors are not particularly helpful for addressing these concerns. Later, we discuss prediction briefly.

VIII. Summarizing the Method and Extending it to Nonselective Institutions to which Students do not Apply

A. Summarizing the Method, its Strengths, and its Frailties

Our value-added method has three parts:

(i) Gathering the results from all of the vertical experiments involving students who are on-the-bubble applicants at some selective school and who are exposed to randomization by admission staff;

(ii) Gathering the results from all of the horizontal experiments involving students who must choose fairly randomly between equally selective colleges that interest them;

(iii) Combining the results of the experiments using the regression dictated by paired comparison theory.

Because we wish to be transparent, let us summarize what we perceive to be the strengths and frailties of this method. The first strength is tackling selection head-on:

this is the issue in estimating colleges' value-added so we have put our remedies for it front and center. The second strength is transparency: we attempt to describe the natural experiments clearly in layperson's language. Of course, the reader maintains his prerogative not to believe that the experiments are sufficiently well-motivated to be credible, but we hope that he will at least understand what experiments generate the results. The third strength is combining the results of all of the experiments efficiently while imposing minimal assumptions in the process. The fourth strength is that the method works in real time with real data, albeit comprehensive, highly accurate data.

We make numerous minor decisions to implement a "base case" for the method. An example is choosing the percentiles for the horizontal experiments. However, we do not view these decisions as frailties but as motivation for robustness testing.

In our view, the main frailty of the method so far concerns the bubble range. It is obvious for selective schools that depend on test scores for applicant triage. However, there is no bubble range for nonselective schools and the range can be non-obvious at schools that are only slightly selective. Of course, we can construct value-added estimates for these nonselective and only slightly schools because students who enroll in them are often on-the-bubble at more selective schools. But, horizontal comparisons for them are shaky because although they do not practice much or any selection, their student bodies may nevertheless differ because they draw from pools that differ on geography or family background. To ensure that value-added is well estimated for nonselective and only slightly selective schools, we extend the method with their circumstances in mind.

B. Extending the Method to Nonselective Schools

In defining student types, we not only grouped students by their test scores but also by the schools to which they applied. We argued that their application behavior was important not only because it revealed their interests but also because it revealed the schools among which they were choosing. Knowing the choice set is especially useful for the horizontal experiments where students are randomizing.

Students who only consider nonselective schools and/or slightly selective schools often do not send their test scores to any college. Since they know with certainty that they will be able to enroll in these schools, they do not really apply. They just fill out paperwork when they register as students. This does not mean, however, that they considered only

the school in which they enrolled. Such students often consider a few local institutions. Because such students' experiences contain valuable information, we would like to create the equivalent of their application portfolios.

To do this, we rely on the fact that when students enroll in nonselective institutions, they are most often very local and are the schools often attended by other students in their high schools. We create a "consideration portfolio" for all students who send scores to no schools or only to local schools that are nonselective or only slightly selective. Each consideration portfolio contains the institution that the student attended plus any institution attended by at least 5 percent of students from the last four graduating classes at the student's high school.[22] Once we have constructed students' consideration portfolios, we treat them exactly as we treat students' application portfolios in all of the foregoing procedures. By doing this, we extend the method so that it incorporates many more horizontal experiments: head-to-head results among nonselective and slightly selective schools that compete for the same local pool of students. The results of these head-to-head local experiments are connected by the paired comparison regression.

C. No versus Any Postsecondary School

Because the U.S. is generously supplied with nonselective postsecondary schools, students are fully able to self-select between high-school-only and nonselective postsecondary schools. Moreover, even the sign of the bias that arises from this self-selection is doubtful. On the one hand, students who are especially motivated may enroll in a nonselective postsecondary school rather than be content with a high school degree or GED. On the other hand, the most competent high school degree holders (among those who are not academically gifted enough for selective college) may obtain jobs. This would imply that students who enroll in nonselective institutions are negatively selected.

It is notoriously hard to find plausible experiments that eliminate selection bias at the no-postsecondary-school versus nonselective-postsecondary-school margin. Our vertical experiments are out of the question because nonselective schools have no bubble. Our horizontal experiments are also out of the question because we have no way to guarantee that a nonselective school is in an indifference set with a job for high school

[22] The four classes include the student's own.

graduate.

For the no-postsecondary-school versus nonselective-postsecondary-school selection problem, we concede that we do not have a remedy that is both credible and usable across all or even most nonselective institutions. Therefore, when we report value-added results, we always normalize the category of postsecondary institutions with the lowest value-added to zero. We caution readers against interpreting all of the positive value-added estimates as positive relative to no-postsecondary-school since the normalized-to-zero institutions may well have negative value relative to no-postsecondary-school. For many questions, a reader may simply ignore the nonselective-versus-no-postsecondary margin. Alternatively, a reader may add to the normalized zero her favorite estimate of the return to nonselective college versus no postsecondary school. (See Oreopoulos and Petronijevic (2013) and Barrow and Malamud (forthcoming) for recent reviews of the evidence.) However, we caution the reader that the most credible estimates from this literature tend to be generated by narrowly defined natural experiments such as the opening of a nonselective college in an area that had no postsecondary institution previously. Such estimates are therefore not representative.

IX. The Value-Added Results

Using the method described above, we estimate value-added for 6,822 U.S. postsecondary institutions. (The missing institutions are recently created or have very small undergraduate enrollment.) We first describe the results that rely on all the experiments including the horizontal experiments based on consideration portfolios.

Table 1 shows value-added estimates for institutions by their selectivity. Each institution is classified by where its 25th percentile scores fit into the national score distribution. For instance, a school is very selective if its own 25th percentile scores are greater than equal to the 90th national percentile. The table shows the mean value-added of institutions in each category. It also shows results for the institutions whose value-added puts them at the 10th and 90th percentiles of value-added within the category. Note that the category of schools with the lowest estimated value-added has their value-added normalized to zero (shaded cell with bold typeface).

The first thing we observe in Table 1 is that more selective U.S. postsecondary

institutions have higher value-added. The relationship is monotonic. This is not altogether surprising: the more selective institutions spend considerably more per student on instruction and related activities. Thus, they must have higher value-added if they are to have any chance of having similar rates of return as less selective schools. Indeed, it is important to recognize that the monotonically positive relationship between value-added and selectivity does not necessary indicate that more selective institutions generate higher rates of return. Because they spend so much more per student and may trigger students to spend more on graduate and professional education as well, a weighing of value-added versus additional costs is needed before we drawn conclusions about rates of return. Later we return to this point.

What is perhaps most striking in Table 1 is that value-added varies so greatly among low selectivity institutions. The 90-10 percentile difference in value-added is wider among the lowest selectivity institutions than the very highest selectivity institutions. Relative to mean value-added, the divergence among low selectivity schools is extremely large. In contrast, schools of middling selectivity do not vary much in value-added within a category.

This evidence suggests that nonselective and only slightly selective schools' value-added varies widely. It is not that all of these schools have low value-added. Rather, even within a group that a student might easily perceive to be comparable, value-added varies dramatically. This result is especially interesting because these are precisely the schools on which can be hardest for a student to obtain accurate information. While some of these schools publish information about their student bodies, graduation rates, net prices, student loan default rates, and post-enrollment outcomes, others do not. Only some of these schools participate in the Common Data Set that is used by college guides such as Barron's and Peterson's.

In short, it is clearly possible for a student who is choosing among equally nonselective schools to end up with considerably higher or lower value-added. That is, it is easy for students to "make mistakes" in this part of the postsecondary market.

Table 2 shows value-added estimates for institutions by their control--private non-profit, public, private for-profit--as well their selectivity. (In order to show estimates by control, the selectivity categories are coarsened substantially relative to Table 1.) The

table shows mean value-added as well as value-added for the institutions at the 10th and 90th percentiles of value-added within each category. The category of schools with the lowest estimated value-added again has value-added normalized to zero (bold typeface).

The most striking pattern in Table 2 is that the for-profit institutions stand out within any given selectivity category. As a rule, the public and non-profit institutions of the same selectivity have similar value-added. It is not merely that their mean value-added is similar: their 10th and 90th percentiles are also fairly similar. In contrast, within each selectivity category, the for-profit institutions have much lower mean value-added. Moreover, within each category, the variation in the for-profits' value-added is very wide relative to their mean. (Compare the same statistic for the public and non-profit schools.) This evidence suggests that the for-profit sector is much more diverse for a student able to gain admittance to schools of a certain productivity. This is not to say that all for-profits offer lower value-added than public or non-profit institutions of comparable selectivity. Rather, the evidence suggests that it is easier for students to make mistakes in the for-profit sector because some of the institutions add much less value than others.

The other result worthy of note in Table 2 is shown in the top two rows. Among very selective colleges (colleges whose 25th percentile student scores at or above the 75th national percentile), there is a long right-hand tail to value-added that occurs only in the non-profit sector. Although the 10th percentile and mean value-added do not differ markedly between public and non-profit schools within this category, their 90th percentile value-added is dramatically different. This is not necessarily evidence of high rates of return at very selective non-profits because the same schools exhibit a long right-hand tail in instructional and related spending (Hoxby 1999). That is, these schools spend so much on students that they could have strikingly high value-added and still have modest rates of the returns. What we have learned is that it is not out of the question that their generous instructional spending earns normal returns: careful rate of return calculations are needed.

Table 3 shows value-added estimates for institutions by their annual core spending per full-time equivalent student. Core spending is the sum of instructional spending, academic support, student services, and institutional support. In other words, it is spending that affects undergraduate students rather than researchers, hospitals, public

service activities and the like. Core spending per student varies widely among U.S. postsecondary institutions. The top category is schools that spend at least $35,000 per student per year. Schools in the bottom category spend less than $5,000 per student per year.

Value-added rises monotonically with schools' core spending. This is at least evidence that rates of return may be solid or better at the higher spending schools. However, as with the previous evidence, we cannot conclude that the value-added is high enough to justify the greater spending. This is especially because students who attend the higher spending schools tend to persist in them, thereby enjoying multiple years of costly education. In contrast, students who attend the lowest spending schools often stay only a year. Thus, their total educational cost (which must be weighed in the balance against the value-added) is low.

As in the previous tables, perhaps the most striking thing in Table 3 is that divergence in value-added among the bottom (low-spending, in this case) institutions. It is not that they all have low value-added. Some would appear to offer very good value. Rather, it is that schools with similar spending have markedly different value-added. This suggests that students choosing among low-spending schools can make substantially better and worse choices.

Because we have been emphasizing the variation in value-added, it is worthwhile reminding the reader that the estimates are not based on sample data. Thus, the variation does not come from sampling error. Rather, the interpretation is that a student, by making apparently small changes in her college choices, could experience very different value-added.

Finally, Table 4 shows how the value-added estimates change as we alter the experiments we use. A few patterns are worth noting. First, using the consideration portfolios makes a difference to value-added estimates, but especially among the least selective schools. This suggests that students who do not actively apply to any selective schools are nevertheless making choices that provide important information about value-added. Second, when we go from using only the vertical experiments to using the horizontal experiments as well, the value-added estimates drop somewhat, especially among schools of middling selectivity. (Compare the two right-most columns.) Recall that

the estimates based on vertical experiments are local to on-the-bubble students and the estimates based on horizontal experiments are local to very-likely-to-be-admitted students. Thus, the contrast between the two right-most columns suggests that schools of middling selectivity generate somewhat higher value-added for their relatively marginal students than their students who have incoming preparation well above the median. One possible explanation is that schools of middling selectivity typically do not have sufficient resources to give each student individual attention. Therefore, students may be pulled toward a median amount of learning with the result that initial high achievers gain less than initial low achievers. In any case, this is interesting evidence that calls for further exploration.

X. Policy Relevance, Extensions, Robustness, and Remaining Issues

In the previous section, we presented value-added estimates that cover most U.S. institutions. This demonstrates that the method is feasible.

Our value-added estimates have many immediate applications. Policies that support higher education tend to be evaluated by their effects on enrollment. To convert these enrollment effects to returns, we need the value-added estimates. The tax-exempt status of colleges and the tax deductibility of charitable contributions to colleges generate important, college-specific tax expenditures. Value-added is what we would weigh against these expenditures. Student loans are currently much debated because default rates and loan volumes are at levels that are historically unprecedented (Looney and Yannelis 2015). Value-added would be crucial to any refinement of the loan program that reduces default or aligns volumes with ability-to-repay. By using value-added, the federal government could potentially spend the same amount on student aid while achieving substantially higher effects on earnings and other outcomes.

A. Extensions

By far the most important limitation of the estimates is that they use only wage and salary earnings as an outcome. We focused on earnings both because they are important for evaluating the benefits of federal tax and spending programs and because they are peculiarly convenient for a primarily methodological study. However, colleges' value-added cannot be fully summarized by their effects on wages. The single most important extension to this study would be adding numerous outcomes beyond wages. We are particularly interested in earnings from non-wage sources, employment, occupation, public service,

inventiveness, family, health, student loan repayment, and charitable giving.

The second most important extension to this study would be examining the causal change in educational costs when a student is induced to attend one college rather than another. This would allow us to make rate of return calculations, for students as private individuals, for the federal and other governments, and for society. The same vertical or horizontal experiments that generate the value-added estimates could be used to generate cost estimates. However, it is important to realize that a student who is induced to attend college A may not merely pay more to college A. She may be triggered by college A to attend professional school, take out a subsidized student loan, take a tax credit for tuition and fees, enjoy tuition that is reduced by state appropriations or private dontations, and so on. All of the costs that arise from her college A choice must be weighed in the balance against the value-added caused by college A. That is, college A's effect includes all of the consequences endogenous to the college A experience. Thus, adding up the educational costs is not a trivial matter: knowing where a person enrolled initially and what she paid is not enough.

B. Robustness Checks

There are many parameters in this paper that can be submitted to robustness checks. We have attempted to conduct checks in priority order according to their likelihood of substantially changing the results. We have already tested (i) altering the probability parameters we use to find the on-the-bubble ranges; (ii) altering the method of fitting enrollment probabilities, again for the bubble range; (iii) changing the percentiles and cushions that we use to define indifference sets of colleges; (iv) using geography alone (not high schools) to construct consideration portfolios; (v) using natural log wage differences. While all of these checks alter the results slightly, they do not change the prominent patterns in the results. Therefore, we defer them for a later paper that focuses on results rather than methodology.

The one robustness check that we have not attempted but interests us greatly is adding family income (or need for financial aid) to the definition of student types. To some extent, we did this by using high schools to define consideration portfolios. However, more refinement is possible. Adding family income to student types would further lessen the chance that students choose between horizontally equal colleges on the basis of income "fit"

or that admissions staff take financial need into account when making admissions decisions. For logistical reasons, though, we would have difficulty refining types on family income unless we simultaneously coarsened types on assessment scores. Thus, this is a robustness check that probably involves a trade-off: the experiments might improve on one dimension but worsen on another.

C. Remaining Issues

Value-added estimates are necessarily backward-looking: they must depend on the experience of previous cohorts of students. Yet, in many applications, we need accurate predictions of value-added for current or future students. In order to generate such predictions, we would need to model how value-added (among past students) was affected by the economy, curriculum, college resources, technology and other factors likely to exercise a major influence on returns to college. Such modeling is beyond the scope of this paper, and prediction is no easier for value-added than for other economic variables. In particular, we expect prediction to be shaky for changes in the macro environment: business cycles, waves of technological innovation, important changes in world trade.

Many people suspect that there are differences in value-added by major or program within colleges. At least for colleges with large enrollment, it would be logistically feasible to estimate value-added by major or program. The difficulty is not the logistics but selection. In the U.S., students choose college majors in a fluid way, often shifting their focus based on their experiences in introductory classes. For instance, a student who struggles in introductory chemistry classes is likely to switch his interest from the pre-medicine major to another. Addressing the resulting selection problem, which probably generates serious bias, has proven to be extremely difficult for researchers. Thus, generating value-added estimates by major is a task for another paper.

In contrast, it would be fairly easy to extend our method to estimate value-added for students with different predetermined (pre-college) preparation, career goals, expressed interest in math versus the humanities, and other characteristics.

XI. References

Avery, Christopher, Mark Glickman, Caroline Hoxby, and Andrew Metrick. 2013. "A

Revealed Preference Ranking of U.S. Colleges and Universities," Quarterly Journal of Economics, 128 (1): 1-45.

Barrow, Lisa, and Ofer Malamud. forthcoming. "Is College a Worthwhile Investment?," Annual Review of Economics.

Bulman, George B. and Caroline M. Hoxby. 2015. "The Returns to the Federal Tax Credits for Higher Education." Tax Policy and the Economy 29: 1-69.

Chetty, Raj, John N. Friedman, and Jonah E. Rockoff. 2014. "Measuring the Impacts of Teachers II: Teacher Value-Added and Student Outcomes in Adulthood," American Economic Review, 104:9. http://www.aeaweb.org/articles.php?doi=10.1257/aer.104.9.2633

Cohodes, Sarah, and Joshua Goodman. 2012. "First Degree Earns: The Impact of College Quality on College Completion Rates," HKS Faculty Research Working Paper Series RWP12-033. http://web.hks.harvard.edu/publications/getFile.aspx?Id=836

Deming, David, Claudia Goldin, and Lawrence Katz. 2011. "The For-Profit Postsecondary School Sector: Nimble Critters or Agile Predators?" NBER Working Paper 17710.

Goodman, Joshua, Michael Hurwitz, Jonathan Smith. 2015. "College Access, Initial College Choice and Degree Completion," NBER Working Paper 20996 http://www.nber.org/papers/w20996

Harville, David A. 2003. "The Selection or Seeding of College Basketball or Football Teams for Postseason Competition," Journal of the American Statistical Association, 98:461: 17-27. http://www.jstor.org/stable/30045190

Hastings, Justine, Christopher Neilson, and Seth Zimmerman. 2012. "Determinants of Causal Returns to Postsecondary Education in Chile: What's Luck Got to Do With It?" NBER conference paper.

Hoekstra, Mark. 2009. “The Effect of Attending the Flagship State University on Earnings: A Discontinuity-Based Approach,” Review of Economics and Statistics, 91 (4): 717-724.

Hoxby, Caroline, and Sarah Turner. 2013. “Expanding College Opportunities for Low-Income, High-Achieving Students,” Stanford Institute for Economic Research Discussion Paper 12-014.

Hoxby, Caroline. 2009. "The Changing Selectivity of American Colleges," Journal of Economic Perspectives, 23(4): 95-118.

Kaufmann, Katja Maria, Matthias Messner, and Alex Solis. 2012. “Returns to Elite Higher Education in the Marriage Market: Evidence from Chile,” Bocconi University working paper. http://tinyurl.com/kaufmanncollrd

Langville, Amy N. and Carl D. Meyer. 2012. Who's #1? The Science of Rating and Ranking. Princeton: Princeton University Press.

Looney, Adam and Constantine Yannelis. 2015. "A Crisis in Student Loans? The Consequences of Non-Traditional Borrowers for Delinquency in the Market," U.S. Treasury and Stanford University typescript.

National Center for Education Statistics, Institute for Education Sciences, U.S. Department of Education. Integrated Postsecondary Education Data System (data as of July 2015 nces.ed.gov website).

Oreopoulos, Philip and Uros Petronijevic. 2013. "Making college worth it: A review of research on the returns to higher education," in The Future of Children: Postsecondary Education in the United States, editors Lisa Barrow, Thomas Brock, Cecelia E. Rouse, 23(1): 41-65.

Saavedra, Juan Estaban. 2009. “The Learning and Early Labor Market Effects of College

Quality: A Regression Discontinuity Analysis." Rand Corporation working paper. http://tinyurl.com/saavedracollrd-pdf

Stern, Steven E. 2011. "Moderated Paired Comparisons: A Generalized Bradley-Terry model for Continuous Data Using a Discontinuous Penalized Likelihood Function," Applied Statistics, 60(3): 397-415.

Zimmerman, Seth D. 2014. "The Returns to College Admission for Academically Marginal Students," Journal of Labor Economics, 32(4): 711-754.

Table 1
Estimated Value-Added of Postsecondary Institutions by their Selectivity

Institutions whose 25th percentile students have math and verbal scores greater than or equal to the...	Mean value-added of institutions	Value-added of 10th %ile institution	Value-added of 90th %ile institution
90th national percentile	90,562	86,381	100,364
85th national percentile (and not listed above)	66,412	50,836	79,952
80th national percentile (and not listed above)	56,258	51,414	62,537
75th national percentile (and not listed above)	46,585	39,995	51,084
70th national percentile (and not listed above)	44,827	38,299	52,133
65th national percentile (and not listed above)	40,018	33,885	43,292
60th national percentile (and not listed above)	37,297	29,365	42,686
55th national percentile (and not listed above)	34,190	29,180	38,314
50th national percentile (and not listed above)	33,088	27,907	38,736
45th national percentile (and not listed above)	29,678	25,682	34,072
40th national percentile (and not listed above)	25,233	20,003	30,252
35th national percentile (and not listed above)	24,530	19,651	28,986
30th national percentile (and not listed above)	21,941	17,168	28,706
25th national percentile (and not listed above)	20,057	16,107	25,129
20th national percentile (and not listed above)	15,819	10,454	21,275
15th national percentile (and not listed above)	10,850	-1,301	20,503
10th national percentile (and not listed above)	8,026	-4,689	18,520
0th national percentile (and not listed above)	0	-14,056	13,585

Notes: The value-added shown in the shaded cell is normalized to zero since we do not attempt to identify how the lowest value-added institutions compare to no college at all. Readers should interpret all other estimates in the table relative to this normalization. Readers are cautioned against interpreting all positive estimates as positive relative to no college at all. See text for further discussion. Source is author's calculations based on type-treatment-level data and institutions' score distributions.

Table 2
Estimated Value-Added of Postsecondary Institutions by their Control and Selectivity

Institutions whose 25th percentile students have math and verbal scores greater than or equal to the...	Control	Mean value-added of institutions	Value-added of 10th %ile institution	Value-added of 90th %ile institution
75th national percentile	non-profit	68,603	51,220	91,736
75th national percentile	public	53,707	51,949	54,525
75th national percentile	for-profit	n/a	n/a	n/a
50th national percentile (and not listed above)	non-profit	42,049	33,408	50,188
50th national percentile (and not listed above)	public	41,312	36,449	45,948
50th national percentile (and not listed above)	for-profit	19,643	-2,823	31,214
25th national percentile (and not listed above)	non-profit	27,413	20,273	35,787
25th national percentile (and not listed above)	public	28,549	23,442	34,028
25th national percentile (and not listed above)	for-profit	7,332	-5,439	25,147
0th national percentile (and not listed above)	non-profit	13,574	2,512	24,422
0th national percentile (and not listed above)	public	14,703	-579	24,774
0th national percentile (and not listed above)	for-profit	0	-27,069	18,504

Notes: The value-added shown in the shaded cell is normalized to zero since we do not attempt to identify how the lowest value-added institutions compare to no college at all. Readers should interpret all other estimates in the table relative to this normalization. Readers are cautioned against interpreting all positive estimates as positive relative to no college at all. See text for further discussion. Source is author's calculations based on type-treatment level data and the Integrated Postsecondary Education Data System (U.S. Department of Education 2015).

Table 3
Estimated Value-Added of Postsecondary Institutions by Core Student-Related Spending

Institutions with annual core spending per full-time equivalent student of ...	Mean value-added of institutions	Value-added of 10th %ile institution	Value-added of 90th %ile institution
$35,000 and up	60,005	39,873	86,938
$30,000 to $34,999	38,950	25,808	55,865
$25,000 to $29,999	36,059	38,037	46,200
$20,000 to $24,999	30,341	20,804	35,558
$15,000 to $19,999	25,083	13,751	37,187
$10,000 to $14,999	19,920	2,855	32,586
$5,000 to $9,999	7,952	-7,503	29,532
$0 to 4,999	0	-16,100	15,905

Notes: The value-added shown in the shaded cell is normalized to zero since we do not attempt to identify how the lowest value-added institutions compare to no college at all. Readers should interpret all other estimates in the table relative to this normalization. Readers are cautioned against interpreting all positive estimates as positive relative to no college at all. See text for further discussion. Core student-related spending of a postsecondary institution is the sum of instructional spending, academic support, student services, and institutional support. It excludes numerous expenditures that are not closely related to the undergraduate student experience. For instance, it excludes research, public service, maintenance and operations, and medical/professional schools. The source is author's calculations based on type-treatment level data and the Integrated Postsecondary Education Data System (U.S. Department of Education 2015).

Table 4
Selectivity of Value-Added Estimates to the Inclusion of Different Experiments

Institutions whose 25th percentile students have math and verbal scores greater than or equal to the...	Mean value-added of institutions		
	using all the experiments	using both vertical and horizontal experiments except those based on consideration portfolios	using only the vertical on-the-bubble experiments
90th national percentile	90,562	89,453	85,139
85th national percentile (and not listed above)	66,412	68,435	69,721
80th national percentile (and not listed above)	56,258	57,865	63,367
75th national percentile (and not listed above)	46,585	51,234	57,887
70th national percentile (and not listed above)	44,827	46.215	52,748
65th national percentile (and not listed above)	40,018	43,758	52,435
60th national percentile (and not listed above)	37,297	41,895	49,978
55th national percentile (and not listed above)	34,190	38,670	49,158
50th national percentile (and not listed above)	33,088	35,852	44,019
45th national percentile (and not listed above)	29,678	33,221	39,511
40th national percentile (and not listed above)	25,233	27,599	34,687
35th national percentile (and not listed above)	24,530	26,431	33,038
30th national percentile (and not listed above)	21,941	24,079	32,774
25th national percentile (and not listed above)	20,057	22,308	28,637
20th national percentile (and not listed above)	15,819	17,384	19,120
15th national percentile (and not listed above)	10,850	14,550	16,383
10th national percentile (and not listed above)	8,026	4,723	4,819
0th national percentile (and not listed above)	0	-328	-287

Notes: The value-added shown in the shaded cell is normalized to zero since we do not attempt to identify how the lowest value-added institutions compare to no college at all. Readers should interpret all other estimates in the table relative to this normalization. Readers are cautioned against interpreting all positive estimates as positive relative to no college at all. See text for further discussion. Source is author's calculations based on type-treatment-level data and institutions' score distributions.